A Bayesian approach for incorporating expert opinions into decision support systems: A case study of online consumer-satisfaction detection

Coussement, K[a], Benoit, DF[b], Antioco, M[c]

[a] *IÉSEG School of Management Universit Catholique de Lille (LEM, UMR CNRS 8179), Department of Marketing, 3 Rue de la Digue, F-59000, Lille, France*
[b] *Ghent University, Faculty of Economics and Business Administration, Tweekerkenstraat 2, B-9000 Ghent, Belgium*
[c] *EDHEC Business School, Department of Marketing, 24 Avenue Gustave Delory, F-59057 Roubaix, France*

Abstract

Interest in the use of (big) company data and data-mining models to guide decisions exploded in recent years. In many domains there are human experts whose knowledge is essential in building, interpreting and applying these models. However, the impact of integrating expert opinions into the decision-making process has not been sufficiently investigated. This research gap deserves attention because the triangulation of information sources is critical for the success of analytical projects. This paper contributes to the decision-making literature by (a) detailing the natural advantages of the Bayesian framework for fusing multiple information sources into one decision support system (DSS), (b) confirming the necessity for adjusted methods in this data-explosion era, and (c) opening the path to future applications of Bayesian DSSs in other organizational research contexts. In concrete, we propose a Bayesian decision support framework that formally fuses subjective human expert opinions with more objective organizational information. We empirically test the proposed Bayesian fusion approach in the context of a customer-satisfaction prediction study and show how it improves the prediction performance of the human experts and a data-mining model ignoring expert information.

Keywords: knowledge fusion, expert system, domain knowledge, classification, Bayes, text mining, location commensurate power prior

1. Introduction

Organizational decision making often relies on a collection of *intangible* capabilities, which are invisible, subjective human-driven phenomena that include organizational routines and employee learning, and *tangible* capabilities in the form of procedural knowledge systems [1]. With the advent of (statistical) data-mining tools and computing power, the tangible capabilities for organisational decision making have become more important. In recent years this process has accelerated as a result of the exponential growth of electronically stored information, which is available to companies, organizations, and individuals. The literature on decision support systems highlights several application domains that have been significantly affected by this trend, including credit risk [2, 3], bankruptcy prediction [4], customer relationship management [5, 6], and fraud detection [7].

Typically, procedural knowledge systems take the form of statistical techniques that are incorporated into a data-mining system (DMS). In line with [8], we define a DMS as the "complete" system — the database or data warehouse, software for mining and analysis, the knowledge derived from these, and the part of the system that supports managerial decision making in a business setting. Traditional DMSs take information about resolved problems and their solutions as input. They then extract rules from that data and use those rules to predict likely outcomes of other cases.

By uncovering patterns or knowledge in the data itself, a DMS obviates the need to elicit knowledge from human experts [9]. To some extent, this is desirable, as expert knowledge might not be available or easily formalized in some industries. In addition, expert knowledge tends to be less objective in nature, and human experts cannot always be relied on to give accurate assessments given their limited reasoning capacities [10]. In contrast, DMSs are labor-saving, intelligent, cognition-based systems that offer the consistency and efficiency that a human expert may lack [11].

However, in many domains, human experts who have developed their expertise over an extended period of time possess important intangible information. Throughout this paper, we refer to such information as an expert system (ES). Despite the existence of well-performing DMSs, human expertise remains dominant in the decision making process [8]. Traditional DMSs do not leave much room for the intangible capabilities of the organisation, as they have not been developed to account for non-data based or subjective information. At present, ESs and DMSs basically remain separate types of information and decision systems. Nevertheless, while findings regarding whether ESs make more accurate and efficient judgments than DMSs are inconclusive [12], one might wonder whether a combination of the two types of systems would result in better overall judgments than a reliance on one or the other.

This paper makes three key contributions to the decision-making literature. First, we provide insight into the natural advantages of the Bayesian philosophy for fusing multiple information sources into one decision support system (DSS). Second, we confirm the necessity of the continuing the search for new or revised methods in this multi-angle information era. Third, we open the path to future applications of Bayesian DSSs in other organizational research contexts. To accomplish the above, we introduce a Bayesian framework that is ideally suited for fusing ESs with DMSs, and we show its beneficial impact on improved decision support. We refer to this fusion of information sources as *fused decision support system* (FDSS). Our approach is based on a blend of recent modeling developments introduced by researchers in the medical [13] and engineering sciences [14]. We empirically benchmark our Bayesian FDSS against an ES and a DMS, both of which use single-source information, in an online setting aimed at detecting consumer satisfaction. This case-study context serves as an ideal test bed because many organizations increasingly focus on online consumer reviews, using web-scraping techniques and advanced big-data analytics to estimate customer satisfaction with the company's product/service strategy. In this regard, organizations rely on internal or external human assessors, who make judgements regarding consumers' satisfaction. The question is

whether this expensive expert-labelling strategy (ES) is effective relative to the more efficient DMSs, which automatically detect the satisfaction level by text mining the content of the consumer reviews, or whether it is crucial for decision makers to fuse all available information sources into a Bayesian FDSS.

The remainder of this paper is structured as follows. First, we revisit ESs and DMSs in organizational decision making, and we discuss how their relative importance has shifted over time. Second, we also describe the quantification of expert opinions and their incorporation into an FDSS. Third, we benchmark the predictive performance of the ES, the DMS, and our novel Bayesian FDSS approach in a real-life case study of customer-satisfaction modeling using online product reviews. Finally, we discuss the implications of using an FDSS for combining human experts' opinions with statistical predictions, before we present our conclusions.

2. Human expert systems, data-mining systems and information fusion: an overview

An expert — the sub-unit of an ES — is someone who has knowledge, and who is capable of efficiently and effectively communicating and using that knowledge during a decision-making process [15]. In the past, human experts often played a monocratic role in managerial decision making. The status of expert is granted on the basis of the individual's professional characteristics and track record, and intuition has been shown to play a critical role in expert decision making [16].

However, experts are not completely free of biases in their judgments of a situation. Potential differences in cognitive styles in terms of what experts think and how they think may lead them to incorrect or biased conclusions [17, 18].

Organizational decision-making processes have undergone a tremendous shift in the last twenty years. DMSs which merely interpret data and make automated decisions regarding the best solution to a problem are becoming very popular. The popularity of DMSs is partially attributable, in part, to the objectivity

and, in part, to the explosion of internal company data collected through recent developments in information technology. The limited information-processing capabilities of human experts means that machines are necessary for coping with the exponential growth in data availability. This massive collection of data, which is often referred to as "big data" [19] offers tremendous opportunities for information systems (IS) researchers and managers, who can incorporate new technologies into DMSs [20]. For instance, in their bibliometric study, Chen et al. [21] report a steep increase in academic publications related to DMSs using big data and business analytics. *Decision Support Systems* ranks as leading IS journal for this type of publication. Moreover, the *IBM Tech Trends Report* [22] identifies business analytics, which is an inherent part of DMSs, as one of the major technology trends of this decade based on a survey of more than 1,200 decision makers from 16 different industries and 13 countries spanning both mature and growth markets.

McAfee and Brynjolfsson [23] estimate that 2.5 billion gigabytes of company data are created every month, and that this number will double every 40 months. The need to effectively and efficiently manage the inflow of company data and the necessity of converting the underlying data patterns into relevant company insights have led to the ever-growing popularity of DMSs. However, despite the initial intention to use DMSs to replicate and replace the decisions of ESs [24], DMSs have limited decision-making abilities because they are best suited for solving problems that have clear boundaries. Moreover, they have limited reasoning capacity [25].

A number of studies look into information fusion for decision support. They find that the inclusion of expert knowledge in DMSs adds value along two dimensions [26, 27]. First, the opinion of the expert is valuable in the independent variable definition phase and when deciding which variables should go into the DMS. In fact, prior studies investigate how domain knowledge helps to define additional, high-level independent variables that are usable by the DMS [28, 29]. Weiss et al. [30] build a DMS using only expert knowledge as input data to predict promising sales leads. Sinha and Zhao [9] combine expert knowledge with

a DMS in the context of credit ratings. To improve the predictive performance of the DMS, these authors ask domain experts to make educated guesses about the credit rating based on characteristics of the applicant and the loan. To test the effects of the inclusion of domain knowledge, the expert predictions are then used as additional inputs for the DMS. The authors find a substantial increase in model performance, which leads them to conclude that fusion of ES and DMS results in substantial monetary benefits for the company. While these studies clearly show the importance of the interplay between an ES and DMS, they limit the role of the domain experts to defining and shaping inputs used in the DMS.

Second, domain experts can contribute to making DMS output consistent with the extant domain knowledge [31]. Previous research primarily operationalizes this aspect by introducing monotonicity constraints on the link between the independent variable(s) and the dependent variable [31]. Therefore, the ES can interact with the DMS during pre-processing [32], classification model building [33] or post-processing [34]. For instance, Lima et al. [33] propose that when the sign of the parameter of an independent variable is the opposite of the sign the expert expects, it should not be incorporated into the analysis. Other authors describe rule-set extraction methods as a valuable approach to bringing domain knowledge into the DMS. See [35] for an overview and comparison of rule-extraction methods in the context of customer churn prediction.

Our Bayesian FDSS method proposes four improvements over the existing methods. First, our method does not only rely on expert opinions in the DSS (e.g., [30]). Moreover it overcomes the drawbacks of Sinha and Zhao [9]'s approach in which domain experts have to evaluate every unit of analysis. For instance, if the goal is to assign a credit score for 1,000 potential lenders, the expert(s) have to manually evaluate each applicant. In many real-life contexts, this procedure is too time consuming and expensive given the large number of cases that have to be screened on a daily basis.

Second, previously used knowledge-fusion approaches impose monotonicity constraints to verify whether the link between the independent variables and

the dependent variable is line with the domain expectations. This is a difficult, yet necessary, exercise [31]. More specifically, this means that these methods assume that: (i) the independent variables are interpretable, (ii) the model parameters/output that serve as the bridge between the independent variables and the dependent variable are intuitive enough for domain experts to interpret, and (iii) the number of independent variables is limited (to make the exercise feasible for the domain expert). Our method simplifies this exercise by incorporating only the experts' opinions on the dependent variable.

Third, building FDSS could improve the predictive performance of the DMS (e.g., [33]). The proposed Bayesian FDSS will rarely perform worse than the corresponding DMS because it relies on the location-commensurate power prior in order to assign suitable weights to the ES and DMS.

Fourth, the Bayesian framework provides a sound theoretical foundation for including subjective information in the analysis. In contrast to frequentist FDSS approaches, it uses the Bayesian prior distribution as the preferred way to integrate expert knowledge into the model [13].

3. Method

3.1. Quantifying expert opinions

One difficulty with combining expert opinions and data-mining models is that the two information sources are often very different in nature. Human expertise does not necessarily relate to the unknown parameters of data-mining algorithms. Consider, for example, a simple linear regression model in which the dependent variable is *customer satisfaction* and the independent variable is *number of complaints*. While most experts would agree that the beta coefficient for *number of complaints* should be negative, the exact magnitude of the beta coefficient might be difficult to assess. This is even more relevant for more advanced data-mining models. For example, the regression coefficients in the logistic regression model represent the effect on the log odds of the event when

the predictor is increased by one unit. Clearly, these parameters are difficult to assess, even for experts.

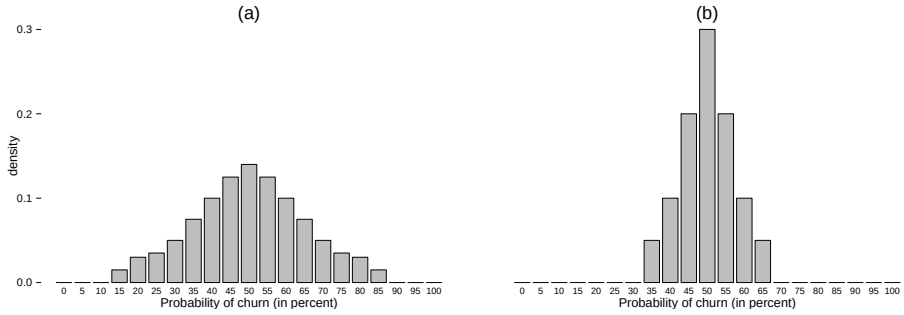

Figure 1: Example of a quantified expert opinion about the probability of churn for a customer that filed one complaint in the last year. The expert opinion in panel (a) is less certain than that in panel (b).

While it is difficult for experts to have an opinion about abstract model parameters, they often have a clear idea about observable quantities. For example, a customer-relationship manager can be asked to give his or her opinion about the probability of churn for a customer who filed one complaint in the last year. Figure 1 provides an example of such an exercise. Panel (a) indicates that the expert believes that this probability is about 50%, and that he or she believes that the probability is highly unlikely to be less than 15% or more than 85%. Panel (b) expresses the opinion of another expert who also believes the likelihood of churn for such a customer is 50%. However, the uncertainty is much lower. This customer-relationship manager is almost 100% certain that the churn probability is between 35% and 65%. The second expert's lower uncertainty results in a more peaked distribution, which is more informative. Notably, when the distribution for the observable quantities is known, the distribution for the abstract model parameters can be easily derived by using the expert information as input data [36].

Expert opinions can be very fine grained (e.g., when a probability is given for every possible outcome, as in Figure 1). However, often only limited information about the experts' opinions is available. This is the case when probability

statements are made about combined outcomes. For example, an expert asked whether sales at a given price point would be between x and y could assign a probability of 50%. Also, only limited expert opinions are available when experts are asked about the expected value of the outcome (i.e., "What is your best guess of expected sales at this price point?"). In this case, the outcome is not a distribution, as in Figure 1, but a single point prediction.

	Coarse	Fine grained
Observable quantities	*Within what range will sales be at this price point?*	*What is the expected value of sales at this price point?*
Abstract model parameters	*Within what range will sales change when the price is increased by one unit?*	*What is the expected value of the change in sales when price is increased by one unit?*

Table 1: Questions that can be asked to retrieve expert knowledge. This information can relate to abstract model parameters or to observable quantities, and it can be detailed or vague.

This discussion demonstrates that expert opinions can be quantified in different ways, as shown in Table 1. One approach is to directly quantify the expert's opinions about the abstract model parameters. Alternatively, one can try to quantify the expert's opinions about observable quantities. In both cases, the quantification can be fine-grained — e.g., a probability for every possible outcome — or coarse — e.g., a single point estimate.

3.2. Incorporating expert opinions into data-mining models: a Bayesian approach

In this paper, we propose a Bayesian fusion approach that overcomes the drawbacks of the other approaches presented in the literature review. The Bayesian machinery is extremely well suited for including non-data based information through the principle of the prior distribution. In Bayesian statistics,

Bayes rule is used to estimate the unknown parameters in a model based on observed data and prior expectations about the parameters [37]. The result is a probability statement about the model parameters or the posterior distribution. Formally, the posterior is given by $p(\theta|D)$, which represents the probability of the unknown parameter(s), θ, given the observed data, D. Bayes rule is used to update the initial beliefs about the model parameters (i.e., before having seen the data) with the insights from the data to form the posterior distribution. In this sense, the posterior is the current belief about the model parameters based on the initial beliefs (i.e., the prior) and the information in the data (i.e., the likelihood). Formally, this idea can be written as:

$$p(\theta|D) \propto p(D|\theta)p(\theta). \tag{1}$$

Essential to the approach used in this paper is how to relate the expert opinions to the prior distribution, $p(\theta)$. It is possible to influence the posterior distribution by defining the functional form of the prior. For example, in a simple linear regression context, we could force a regression parameter to have a positive effect on the outcome by defining a prior distribution that is uniform on the positive real line. This would be the Bayesian equivalent of the sign-restriction method [38, 39, 33]. However, more advanced Bayesian alternatives are possible that do not share the drawbacks outlined in Section 2.

Previous research argues that experts' point predictions can be viewed as historical outcomes [40, 41]. By using the covariate matrix of the observed dataset, we can construct a pseudo dataset containing the expert data that represents the expert opinions as if they were historical observations. For the remainder of this paper, we refer to this dataset as *expert data*, D_0, and use *observed data*, D, for the typical data used in the data-mining model.

The most straightforward way of incorporating this expert information is to set the prior distribution $p(\theta)$ in Equation 1 equal to the posterior distribution resulting from the expert data $p(\theta|D_0)$. Formally:

$$p(\theta|D) \propto p(D|\theta)p(\theta|D_0), \qquad \text{where} \qquad p(\theta|D_0) \propto p(D_0|\theta)p(\theta). \tag{2}$$

To avoid giving the same weight to the expert data and the observed data, [40] proposes raising the likelihood of the historical data to the power of $\alpha \in [0, 1]$. This *power prior* approach controls the heaviness of the tails of $p(D_0|\theta)$ and, ultimately, the posterior distribution $p(\theta|D)$.

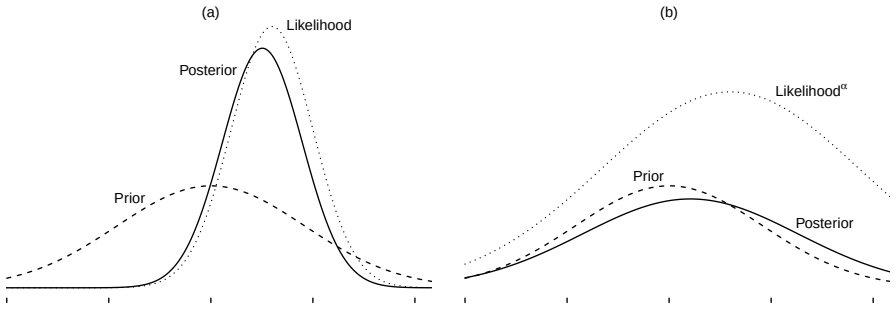

Figure 2: Influence of the power prior on the posterior distribution. Panel (a) full borrowing with $\alpha = 1$ versus panel (b) minimal influence with $\alpha = 0.1$.

This mechanism is illustrated in Figure 2. In panel (a), the likelihood is unchanged (equivalent to $\alpha = 1$) and the posterior is heavily influenced by the likelihood. In panel (b), the likelihood is raised to the power $\alpha = 0.1$. As a result, the posterior is hardly influenced by the likelihood. However, in most situations it is difficult to a priori determine the weight of expert data relative to observed data because of uncertainty regarding the degree to which the expert data agrees with the observed data. The Bayesian solution to this uncertainty is to simply put a prior on α, $p(\alpha)$, and estimate this quantity together with the other unknowns. This extension makes the approach considerably more difficult. The *location commensurate power (LCP) prior* deals with this added complexity [13, 42].

Consider two different parameters: θ_0 for the expert data D_0 and θ for the observed data D. Then define an initial vague or flat prior $p(\theta_0)$, and construct a normal prior on θ with a mean equal to θ_0 and precision (i.e., inverse variance) of τ. The precision τ can be interpreted as a measure of commensurability between θ_0 and θ, and it is used to guide the prior on α. By defining a prior on

τ and normalizing with respect to θ_0, the LCP prior can be completed:

$$p(\theta, \alpha, \tau | D_0) \propto \int \frac{[p(\theta_0|D_0)]^\alpha}{\int [p(\theta_0|D_0)]^\alpha \, d\theta_0} \times N(\theta|\theta_0, \tau^{-1}) \, d\theta_0 \times Be(\alpha|g(\tau), 1) \times p(\tau), \tag{3}$$

where $g(\tau) > 0$ is a function of the commensurability parameter that is small for τ close to 0 and large for large values of τ. We use the suggestion found in [13, 42] in that we set $g(\tau) = \max(\log(\tau), 1)$ to avoid numerical problems with extremely large or small values for τ. Furthermore, $Be(\cdot|\alpha, \beta)$ denotes the beta distribution with parameters α and β.

The intuitive interpretation is as follows. When the expert data and the observed data disagree, the influence of the expert data on the resulting posterior will be minimal because small α values are favored. When the expert data and the observed data agree, the opposite occurs. The advantage of this LCP prior approach is that it balances the amount of strength to borrow from the expert data with statistical integrity [13]. In other words, the method assures that expert knowledge that is inconsistent with the accumulating observed data will have minimal influence on the resulting posterior. As such, in contrast to other models that include expert opinions [**?**], this subtle weighting approach has the advantage that the extended model will rarely do worse than the observed data-only model. In addition, the LCP prior incorporates expert knowledge without requiring the expert to evaluate every case in the dataset, as in [9].

4. Incorporating expert opinions for online consumer-satisfaction detection

4.1. Context

Organizations increasingly rely on big data covering online customer activities to estimate the level of customer satisfaction. In fact, companies are flooded with billions of Internet-based conversations about their products. Online reviews constitute a significant portion of these customer-company interactions [43]. This popular form of word-of-mouth communication takes place on websites such as Amazon.com, Epinions.com, and Reviewcentre.com.

The process of detecting the level of satisfaction indicated in online product reviews which typically focusses on analyzing their content, is an important task for many companies. On the one hand, extremely negative word-of-mouth is generally harmful for companies, as it produces unfavorable effects on various company indicators, such as customer loyalty, and thereby results in increased defection rates, higher customer-acquisition costs [44, 45], and a lower likelihood of repurchases. It also negatively affects consumer behavior, the corporate image, institutional legitimacy, and stakeholders' trust in the organization [46]. On the other hand, extremely positive feedback creates positive snowball effects in which customers are positively influenced by the opinions of current or past users.

Many companies rely on in-house processes and/or specialized services, such as CrowdFlower.com or Amazon MTurk, in which human experts assess or estimate the satisfaction level implied in online product reviews. This gives rise to the question of whether this expert-labeling approach results in accurate and efficient classifications of customer satisfaction.

4.2. An FDSS for consumer-satisfaction detection

Figure 3 shows the process of predicting customer satisfaction from online product reviews using three approaches available to organizations: ES through human experts, DMS, and our novel Bayesian FDSS. As shown in Figure 3, the process of categorizing scraped consumer reviews into multiple satisfaction categories can be broken into four stages: data collection, data preparation, data analysis, and decision and evaluation.

4.2.1. Data collection

We gathered written customer reviews and corresponding satisfaction ratings from an online review website, Reviewcentre.com. Although this website collects reviews across a broad spectrum of product and service categories, we concentrated on reviews of recently launched technology products. As customers commonly give and seek extensive feedback about technology products and ser-

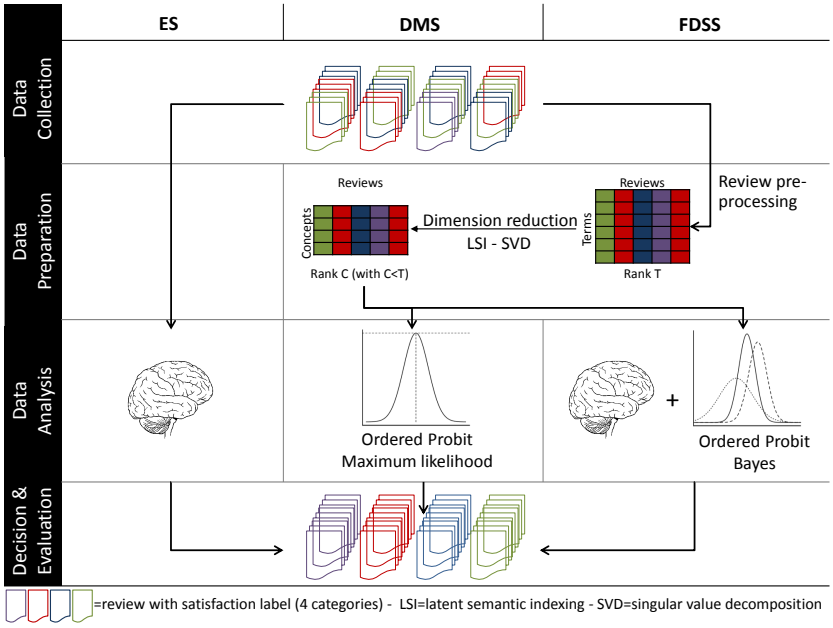

Figure 3: Graphical representation of FDSS, ES, and DMS.

vices [47]. More specifically, we gathered 1,014 customer reviews on technology products, including laptops, portable audio players, and software programs. Within this category, we selected the three most reviewed subcategories and collected all customer reviews in those subcategories as well as information on their satisfaction ratings (range from 0 till 10). The content of these reviews was used in the DMS and FDSS to reveal the customer-satisfaction levels.

The reviews in our data set were also assessed by 507 marketing experts, all of whom were all native English speakers with at least three years of professional experience. They exhibited acceptable inter-rater reliability (AC1 = .591; $p < .001$) [48]. Each expert read ten randomly assigned reviews and estimated each customer's satisfaction on a scale from 0 to 10. All reviews were rated by five experts, which enabled us to obtain overall satisfaction measures that were intended to counterbalance any outliers.

4.2.2. Data preparation

Reviews are textual information sources that have to be transformed into numeric data that can be used in a DMS or an FDSS. In this regard, we followed the guidelines set by Feldman and Sanger [49]. We used the bag-of-words methodology to convert the textual information into numeric representations, while latent semantic indexing (LSI) helped us construct a low-dimensional, concept-by-review matrix.

Pre-processing

We first removed special characters and punctuation from terms. The stream of characters in the review was then converted into a stream of terms or tokens, which is known as tokenization. Special characters, numbers, and punctuation were removed from the text and all terms were converted to lowercase. The terms were then benchmarked to the WordNet database as a first quality check. WordNet is a lexical reference system that holds definitions and semantic relations between words for over 100,000 English terms [50].

At this stage, all distinct terms in the review corpus are independent variables that characterize the content of the review corpus. This results in a high-dimensional, term-by-review matrix that contains thousands of distinct terms and, thus, independent variables. As a term-by-review matrix is not workable from a choice-modeling perspective, we applied several term-filtering practices to reduce the number of terms in the matrix. First, we removed rare terms from the term list, as such terms do not help in future document description. Furthermore, stop words — words that are extremely common (e.g., "a" and "the") — were eliminated. These words appear so frequently in the text that they do not have any discrimination power.

In the next step, term variations were conflated into a single representative form, which is referred to as the "stem". This helps reduce the dimensionality of the term-by-review matrix. An example is the word "inspect," which is the stem for the variants "inspected," "inspecting," "inspection," and "inspections." We used the Snowball Stemmer, which is the most well-known, affix-removal

stemmer [51].

A last step in the review pre-processing phase is the weighting of the terms in accordance with their importance for characterizing the review, and their discrimination power in the total review corpus. Spärck Jones [52] showed significant improvements in performance when using weighted-term vectors. In practice, for each cell in the term-by-review matrix, we converted the raw frequency of the appearance of term t in review r using the weight frequency of term t in review $r(w_{t,r})$:

$$w_{t,r} = tf_{t,r} \times idf_t, \tag{4}$$

with $tf_{t,r}$ equal to the term frequency of term t in review r and idf_t equal to the inverse document frequency of term t, and defined as:

$$idf_t = \ln\left(\frac{N}{df_t}\right). \tag{5}$$

N is equal to the total number of reviews in the corpus, while df_t equals the number of reviews in which term t is present. The term frequency characterizes the representation power of a term for a particular review. The more a term is present in a review, the more important that term is for characterizing the content of that review. The term frequency is counterbalanced by the inverse document frequency, such that the more a term is present in the review corpus, the less discriminating that term is and the lower its weight.

Dimension reduction

The result of the review pre-processing is a high-dimensional, weighted, term-by-review matrix. As this matrix is too large and sparse to be used for predictor variables, dimensionality reduction is necessary. This process is referred to as LSI, as it assumes a latent structure in the review corpus. To reduce the dimensions of the term-by-review matrix, we used SVD [53], which relies on the presence of certain terms in similar reviews to establish relationships between the terms. The SVD method projects reviews from a high-dimensional term

space into an orthonormal, semantic, latent subspace. To decide on the ideal number of dimensions to use to summarize the content of the original term-by-document, we relied on the profile log-likelihood [54]. The final concept-by-review matrix stores the products reviews in rows, while the distinct concept variables are given in the columns. The latter variables are used as *independent variables* in the DMS and FDSS, while customer satisfaction is the *dependent variable*. In addition to determining whether a product fails the customer-usage test (i.e., a satisfaction score of less than 5), we categorized the product reviews into four distinct satisfaction categories, which are described in Table 2.

Description	Rating range	Percentage of reviews
Very unsatisfied	0–2	38.86%
Below par	3–5	12.62%
Above par	6–8	19.62%
Very satisfied	9–10	28.90%

Table 2: Description of satisfaction categories

4.2.3. Data analysis

This study uses the ordered probit model as the baseline choice model in the DMS and FDSS to classify the satisfaction level based on each reviews' content. The ordered probit model is a direct extension of the well-known binary probit regression model. However, the ordered probit model allows for more than two choice options as long as the ordering of the choice options is reasonable, as is clearly the case in our satisfaction-detection setting. The main advantage of modeling the unobserved utilities with a normal distribution is that the model parameters can be estimated with an efficient Gibbs sampling algorithm. Moreover, the normal distribution is one of the most common distributions — together with the logistic distribution — for modelling ordered responses [55].

Define the data $D = \{y_i, X_i\}$, where y_i can take on values $y_i = \{1, \ldots, J\}$, where J is the total number of ordered alternatives (i.e., $J = 4$ in this application) and i is the indicator of the choice maker. X is the matrix of predictor

variables. Also define an unobserved latent variable $y_i^\star$, which represents the unobserved continuous valuation of the choices, so that:

$$y_i = j \quad \text{if} \quad \gamma_j < y_i^\star \leq \gamma_{j+1}, \tag{6}$$

where $\gamma = \{\gamma_1, \ldots, \gamma_{J+1}\}$ is a vector of parameters with $\gamma_1 \leq \cdots \leq \gamma_{J+1}$, $\gamma_1 = -\infty$, $\gamma_2 = 0$ and $\gamma_{J+1} = \infty$. The assumption of standard normal distributed latent variables $y_i^\star$ leads to the following model [56]:

$$Pr(y_i = j) = \Phi(\gamma_{j+1} - x_i'\beta) - \Phi(\gamma_j - x_i'\beta), \tag{7}$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution.

The only unknowns in this model are the regression coefficients given by the vector β. As in all Bayesian analyses, the unknowns need a prior distribution. In this case, we use the LCP prior given by Equation 3. The resulting posterior distribution is analytically intractable but can be sampled using Markov Chain Monte Carlo methods [56]. For the Bayesian FDSS proposed in this paper, the ordered probit model is extended with the LCP prior. An efficient Gibbs sampler was developed to sample from the posterior distribution (see [13, 42] for details).

4.2.4. Decision and evaluation

The predictive performance of the approaches are evaluated using the micro-averaged F1 measure [57]:

$$F1 = \frac{2(prec \times rec)}{prec + rec}, \tag{8}$$

with *prec* denoting precision and defined as:

$$prec = \frac{\sum_{c=1}^{4} TP_c}{\sum_{c=1}^{4} TP_c + FP_c}, \tag{9}$$

and *rec* denoting recall and defined as:

$$rec = \frac{\sum_{c=1}^{4} TP_c}{\sum_{c=1}^{4} TP_c + FN_c}. \tag{10}$$

TP_c is defined as the number of reviews belonging to the satisfaction category c and correctly classified by the algorithm. FP_c is defined as the number of reviews wrongly assigned to the satisfaction category c, while FN_c is defined as the number of reviews assigned incorrectly to satisfaction category c.

Intuitively, the ideal DMS seeks to achieve efficiency (high recall) and effectiveness by not assigning too many reviews to the wrong satisfaction category (high precision). However, as the relationship between precision and recall is negative, a trade-off exists between the two [58]. As both measures are equally important, text-classification studies combine precision and recall measures into the F1 measure. The higher the F1 measure, the better the approach is in terms of detecting the customer-satisfaction level. As the F1 measure ignores true negatives and as its magnitude is mostly determined by the number of true positives, the large satisfaction groups (i.e., "very unsatisfied" and "very satisfied",) dominate the small groups in the global evaluation measure. This is necessary, as an extreme satisfaction level, regardless of whether it is positive or negative, has the biggest impact on company performance.

5. Results

5.1. Experimental setup

To truly evaluate the performance of a choice model, two types of data sets must be created: a training set and a validation set. The training set is used for model estimation, while the estimated choice model is applied to the validation set that contains reviews not used in the estimation phase. The predictive performance of the choice model should be measured using unseen validation data because this enables the researcher to avoid finding relations between the LSI concepts that are used as independent variables and customer satisfaction. Such relations would reflect idiosyncratic characteristics of the training data that do not hold up in the real world. The use of these two data sets mimics a real-life setting in which a choice model used in a predictive context is built on

a review set containing the customer ratings and is then applied to gain insight into the satisfaction level in unlabeled reviews.

However, a single random split of the review data set into a training set and a validation set is known to produce validation set results that are highly dependent on the split. This is a result of sampling variation across possible splits. To avoid such dependencies, we applied our methodological framework 150 times by randomly splitting the original review base into 60% for training and the remaining 40% for validation.

5.2. Comparing predictive performance

Table 3 provides a summary of the F1 performance measures for the satisfaction labeling of the online reviews by ES, the frequentist DMS that only uses the content of the online reviews, and the Bayesian FDSS that aggregates the average decision of the five experts as informative prior with the content of the product reviews. Given the non-normality of the expert system's performance measure, we analyzed the results using Friedman's χ^2-test, which is the non-parametric equivalent of the repeated-measures ANOVA.

ES	DMS	FDSS
.51 (.02)	.48 (.02)	.53 (.02)

Table 3: Overview of the cross-validated F1 measures for the ES, DMS and FDSS. The table shows the mean F1 measures with their standard deviations in parentheses.

The results indicate that significant performance differences exist among the three approaches (Friedman's χ^2-test = 79.39; df = 56; $p < .05$). Furthermore, we used a Nemenyi post-hoc test to mutually compare the approaches. More specifically, we ranked the three approaches for each of the 150 bootstrap samples as follows: the best-performing approach was assigned a rank of three, the second-best approach was assigned a rank of two, and the worst-performing approach was assigned a rank of one. The performance of two approaches is significantly different if the corresponding average ranks differ by at least the

critical difference (CD) with:

$$CD = q_\alpha \sqrt{\frac{k(k+1)}{6M}}, \tag{11}$$

where k equals the total number of approaches to compare, M equals the number of bootstrap samples, and q_α is the critical value, which is based on the studentized range statistic divided by $\sqrt{2}$. In our setting, $k = 3$, $M = 150$, and $q_\alpha = 2.34$ at a 95% confidence level. This means that two approaches are significantly different at the 95% confidence level if the average ranks differ by at least 0.27 (i.e., the CD). Figure 4 summarizes the cross-validated average ranks of the ES, the DMS, and the FDSS. Note that higher ranks indicate better performance. Significant (with $\alpha = 0.05$) differences in performance between approaches are connected.

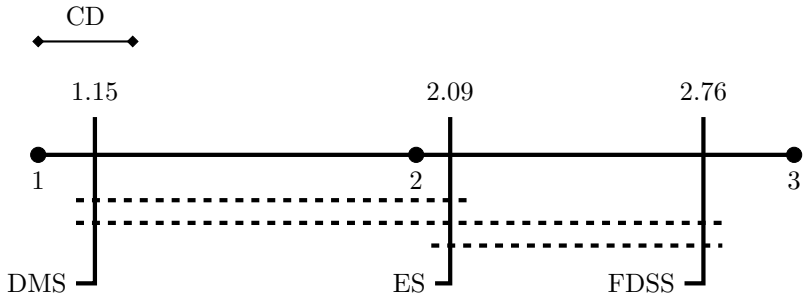

Figure 4: Nemenyi post-hoc test results for the human experts (ES), the data-mining system (DMS), and the newly proposed fused decision support system (FDSS). Dashed lines indicate significant ($p < .05$) differences.

The Nemenyi post-hoc test confirms that the data only approach, DMS, is significantly worse in identifying satisfaction levels than the human experts (ES) and the fused information approach (FDSS). The results also show that the integration of the ES and traditional data-mining approaches leads to a significant improvement in model performance relative to the predictive performance of human experts' predictions or purely data-based models on their own.

6. Discussion

This paper started from the observation that two major sources of information for decision making exist, that is managerial information and data-based information. We argued that both are valid and relevant, but also that both approaches are not easily combined in a single framework. The proposed methodology in this paper is an attempt to solve this problem.

The results show that fusion of information sources clearly achieves better performance than a ES or a DMS. This implies that organizations should not rely solely on quantifiable and less expensive (owing to the high degree of automation) big data, or on the expensive and subjective opinions of human decision makers. Rather, a combination of both information sources helps to ensure higher-quality decisions.

This finding corroborates with conclusions drawn by [30, 33, 9]. However, the methodologies put forward by these authors typically require very time-consuming and demanding information-extraction tasks from the experts. For example, the approach in [9] requires experts to evaluate and give input for every case that needs a prediction.

As the specific nature of the Bayesian methodology proposed in this paper alleviates this, the method is better suited for large-scale applications. Also this makes the proposed approach easier and cheaper to implement than existing fusing methods.

Moreover, the Bayesian approach allows for inclusion of expert knowledge in complex data-mining algorithms. Some existing approaches, such as [**?**] only work for relatively simple data-mining algorithms like linear regression, because experts have to express their knowledge about the model parameters. In the proposed method, expert information is not required to relate to the model or its parameters but only to observable quantities to be predicted. This makes the extraction of expert information less demanding for the experts involved, and also this expert information can be fed to more advanced data-mining algorithms.

7. Conclusion

This article highlights the value of combining rather subjective human expert knowledge with more objective organizational information in DSSs. The idea of information fusion is managerially relevant for two reasons. First, a DSS that does not fully exploit all available information and, thus neglects relevant information like experts' opinions is unlikely to lead to optimal decisions. Second, the opinions of experts, who are often the end users of a DSS, must be heard during the building process if a DSS is to be successfully implemented in a company. DSSs that explicitly take their experts' knowledge into account ensure that this need is met.

In this study, we propose an FDSS that uses the Bayesian philosophy, and we contrast its prediction performance against the single-source benchmarks of a traditional DMS and the opinions of human experts. In contrast to the non-Bayesian methods, which are often at the core of data-mining approaches and theoretically do not allow for the inclusion of this type of subjective information, the incorporation of expert information in the model is natural in the Bayesian framework. More specifically, the Bayesian prior distribution makes it possible to include this additional information in the analysis. Bayes rule then makes sure that the output, or posterior distribution, is a correct representation of the combined prior information and the observed data. When the expert opinions are consistent with the observed data, the posterior distribution will be more peaked and, therefore, be more informative.

This study contributes to the decision-making literature in several ways. First, it creates awareness of the natural advantage of the Bayesian philosophy over the traditional, frequentist, approaches to integrating multiple information sources into one decision-support framework. Second, it stressing the need for adapted tools in this data-explosion era. Third, it highlights the possibility of researching future applications of Bayesian DSSs in other organizational contexts.

We also identify several avenues for future research. While we have focussed

on the Bayesian method for fusing information sources, more research is needed on the types of expert knowledge that influence the performance of the FDSS. The value added by information fusion may not be equally high for the different decision problems a company faces, such as satisfaction prediction, human resource analytics, cross-selling, and customer acquisition. Moreover, the added value of information fusion could vary across industries, requiring additional benchmark studies. Finally, a valuable avenue for further research is how other baseline models can be extended and improved with expert knowledge, and how those new models compare to the benchmark algorithms.

References

[1] A. A. Lado, M. J. Zhang, Expert systems, knowledge development and utilization, and sustained competitive advantage: A resource-based model, Journal of Management 24 (4) (1998) 489–509.

[2] P. Hájek, Municipal credit rating modelling by neural networks, Decision Support Systems 51 (1) (2011) 108–118.

[3] M. C. So, L. C. Thomas, H.-V. Seow, C. Mues, Using a transactor/revolver scorecard to make credit and pricing decisions, Decision Support Systems 59 (0) (2014) 143–151.

[4] D. L. Olson, D. Delen, Y. Meng, Comparative analysis of data mining methods for bankruptcy prediction, Decision Support Systems 52 (2) (2012) 464–473.

[5] S. Moro, P. Cortez, P. Rita, A data-driven approach to predict the success of bank telemarketing, Decision Support Systems 62 (0) (2014) 22–31.

[6] T. Verbraken, F. Goethals, W. Verbeke, B. Baesens, Predicting online channel acceptance with social network data, Decision Support Systems 63 (0) (2014) 104–114.

[7] S. Bhattacharyya, S. Jha, K. Tharakunnel, J. C. Westland, Data mining for credit card fraud: A comparative study, Decision Support Systems 50 (3) (2011) 602–613.

[8] H. Daniels, A. Feelders, M. Velikova, Integrating economic knowledge in data mining algorithms, in: Proceedings of the 8th International Conference on Society for Computational Economics: Computing in Economics and Finance, Aix-en-Provence, 2002.

[9] A. Sinha, H. Zhao, Incorporating domain knowledge into data mining classifiers: An application in direct lending, Decision Support Systems 46 (1) (2008) 287–299.

[10] M. Hilbert, Toward a synthesis of cognitive biases: How noisy information processing can bias human decision making, Psychological Bulletin 138 (2) (2012) 211–237.

[11] D. Williamson, I. Bejar, A. Hone, 'mental model' comparison of automated and human scoring, Journal of Educational Measurement 36 (2) (1999) 158–184.

[12] T. C. Powell, A. DentMicallef, Information technology as competitive advantage: The role of human, business, and technology resources, Strategic Management Journal 18 (5) (1997) 375–405.

[13] B. P. Hobbs, B. P. Carlin, S. J. Mandrekar, D. J. Sargent, Hierarchical commensurate and power prior models for adaptive incorporation of historical information in clinical trials, Biometrics 67 (3) (2011) 1047–1056.

[14] A. Zuashkiani, D. Banjevic, A. K. S. Jardine, Estimating parameters of proportional hazards model based on expert knowledge and statistical data, Journal of the Operational Research Society 60 (12) (2009) 1621–1636.

[15] A. Hart, Knowledge acquisition for expert systems, McGraw-Hill, Inc., 1986.

[16] K. Peters, K. Daniels, G. P. Hodgkinson, S. A. Haslam, Experts judgments of management journal quality: An identity concerns model, Journal of Management.

[17] R. Cooke, Experts in Uncertainty : Opinion and Subjective Probability in Science: Opinion and Subjective Probability in Science, Oxford University Press, USA, 1991.

[18] P. Tetlock, Expert Political Judgment: How Good is It? How Can We Know?, Princeton University Press, 2005.

[19] D. Dean, S. DiGrande, D. Field, A. Lundmark, J. O'Day, P. J., Z. P., The internet economy in the g-20, Boston Consulting Group perspectives.

[20] R. M. Chang, R. J. Kauffman, Y. Kwon, Understanding the paradigm shift to computational social science in the presence of big data, Decision Support Systems 63 (SI) (2014) 67–80.

[21] H. Chen, R. Chiang, V. Storey, Predicting online channel acceptance with social network data, MIS Quarterly 36 (4) (2012) 1165–1188.

[22] I. T. T. Report, The 2012 ibm tech trends report: Fast track to the future (2012).
URL `http://www.ibm.com/developerworks/techntrendsreport`

[23] A. McAfee, E. Brynjolfsson, Big data: The management revolution, Harvard Business Review 90 (10) (2012) 60–+.

[24] T. Hermann, K. Just, Experts' systems instead of expert systems, AI and Society 9 (4) (1995) 321–355.

[25] E. Turban, P. Watkins, Integrating expert systems and decision support systems, Mis Quarterly 10 (2) (1986) 121–136.

[26] I. Kopanas, N. Avouris, S. Daskalaki, The role of domain knowledge in a large scale data mining project, in: I. Vlahavas, C. Spyropoulos (Eds.),

Methods and Applications of Artificial Intelligence, Vol. 2308 of Lecture Notes in Computer Science, Springer Berlin Heidelberg, 2002, pp. 288–299.

[27] F. Alonso, J. Caraça-Valente, A. González, C. Montes, Combining expert knowledge and data mining in a medical diagnosis domain, Expert Systems with Applications 4 (23) (2002) 367–375.

[28] H. Hirsch, M. Noordewier, Using background knowledge to improve inductive learning: a case study in molecular biology, IEEE Expert (10) (1994) 3–6.

[29] R. Ambrosino, B. Buchanan, The use of physician domain knowledge to improve the learning of rule-based models for decision-support, in: Proceedings of the Annual Fall Symposium of the American Medical Informatics Association, Washington DC, 1999, pp. 3–6.

[30] S. Weiss, S. Buckley, S. Kapoor, S. Damgaard, Knowledge-based data mining, in: Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington DC, 2003, pp. 456–461.

[31] D. Martens, J. Vanthienen, W. Verbeke, B. Baesens, Performance of classification models from a user perspective, Decision Support Systems 51 (2011) 782–793.

[32] H. A. Daniels, M. V. Velikova, Derivation of monotone decision models from noisy data, Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on 36 (5) (2006) 705–710.

[33] E. Lima, C. Mues, B. Baesens, Domain knowledge integration in data mining using decision tables: case studies in churn prediction, Journal of the Operational Research Society 60 (8) (2009) 1096–1106.

[34] A. Feelders, M. Pardoel, Pruning for monotone classification trees, in: M. Berthold, H. Lenz, E. Bradley, R. Kruse, C. Borgelt (Eds.), Advances

in Intelligent Data Analysis, Vol. V, Springer Berlin Heidelberg, 2003, pp. 1–12.

[35] W. Verbeke, D. Martens, C. Mues, B. Baesens, Building comprehensible customer churn prediction models with advanced rule induction techniques, Expert Systems with Applications 38 (2011) 2354–2364.

[36] B. Kraan, T. Bedford, Probabilistic inversion of expert judgments in the quantification of model uncertainty, Management Science 51 (6) (2005) 995–1006.

[37] J. Bernardo, A. Smith, Bayesian Theory, John Wiley & Sons, New York, 1994.

[38] H. Daniëls, A. Feelders, Integrating economic knowledge in data mining algorithms, Discussion Paper 2001–63, Tilburg University, Center for Economic Research (2001).

[39] T. van Gestel, D. Martens, B. Baesens, D. Feremans, J. Huysmans, J. Vanthienen, Forecasting and analyzing insurance companies' ratings, International Journal of Forecasting 23 (3) (2007) 513–529.

[40] J. G. Ibrahim, M. H. Chen, Power prior distributions for regression models, Statistical Science 15 (1) (2000) 46–60.

[41] A. Zellner, On assessing prior distributions and bayesian regression analysis with g-prior distributions, in: P. Goel, A. Zellner (Eds.), Studies in Bayesian Econometrics and Statistics, Elsevier Science Publishers, Amsterdam, 1986.

[42] B. P. Hobbs, D. J. Sargent, B. P. Carlin, Commensurate priors for incorporating historical information in clinical trials using general and generalized linear models, Bayesian Analysis 7 (3) (2012) 639–673.

[43] D. Godes, D. Mayzlin, Y. B. Chen, S. Das, C. Dellarocas, B. Pfeiffer, B. Libai, S. Sen, M. Z. Shi, P. Verlegh, The firm's management of social interactions, Marketing Letters 16 (3–4) (2005) 415–428.

[44] S. Gupta, V. Zeithaml, Customer metrics and their impact on financial performance, Marketing Science 25 (6) (2006) 718–739.

[45] R. T. Rust, K. N. Lemon, V. A. Zeithaml, Return on marketing: Using customer equity to focus marketing strategy, Journal of Marketing 68 (1) (2004) 109–127.

[46] X. M. Luo, C. B. Bhattacharya, Corporate social responsibility, customer satisfaction, and market value, Journal of Marketing 70 (4) (2006) 1–18.

[47] F. Zhu, X. Zhang, Impact of online consumer reviews on sales: The moderating role of product and consumer characteristics, Journal of Marketing 74 (2) (2010) 133–148.

[48] K. Gwet, Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Multiple Raters, Advanced Analytics, LLC, 2010.

[49] R. Feldman, J. Sanger, The text mining handbook: advanced approaches in analyzing unstructured data, Cambridge University Press, 2007.

[50] P. University, Wordnet: A lexical database for english.
URL `http://wordnet.princeton.edu`

[51] M. Porter, Snowball stemmer (2001).

[52] K. Spärck Jones, A statistical interpretation of term specificity and its application in retrieval, Taylor Graham Publishing, 1988, pp. 132–142.

[53] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, R. Harshman, Indexing by latent semantic analysis, Journal of the American Society for Information Science 41 (6) (1990) 391–407.

[54] M. Zhu, A. Ghodsi, Automatic dimensionality selection from the scree plot via the use of profile likelihood, Computational Statistics & Data Analysis 51 (2) (2006) 918–930.

[55] W. Greene, Econometric Analysis, Prentice Hall, Upper Saddle River, 2003.

[56] G. Koop, Bayesian econometrics, J. Wiley, 2003.

[57] C. J. Van Rijsbergen, Information retrieval, Butterworths, 1979.

[58] M. Buckland, F. Gey, The relationship between precision and recall, Journal of the American Society for Information Science 45 (1994) 12–19.